BoNC: Bag of N-Characters Model for Word Level Language Identification

Shimaa Ismail[1] and **Mai K. Gallab**[2] and **Hamada Nayel**[2]
[1]Department of Information Systems
[2]Department of Computer Science
Faculty of Computers and Artificial Intelligence
Benha University, Egypt
{shimaa.mustafa, mai.gallab, hamada.ali}@fci.bu.edu.eg

Abstract

This paper describes the model submitted by NLP_BFCAI team for Kanglish shared task held at ICON 2022. The proposed model used a very simple approach based on the word representation. Simple machine learning classification algorithms, Random Forests, Support Vector Machines, Stochastic Gradient Descent and Multi-Layer Perceptron have been implemented. Our submission, RF, securely ranked fifth among all other submissions.

1 Introduction

Language Identification ("LI") is the process of identifying the natural language that a document or a portion of it is written in (Li et al., 2013). A human reader who is familiar with a language may easily recognize material written in that language. Therefore, the goal of LI research is to imitate the capacity of humans to identify these languages. The early attempts to solve this problem were made at the beginning of the century (Tomokiyo and Jones, 2001; Jarvis and Crossley, 2012). After that, there are several computer methods is being used without the assistance of a human using specifically created algorithms and indexing structures. Over the years, LI research has developed methods to recognize human languages. LI is applicable for all forms of information storage that incorporate language, whether digital or not, and applies to every modality of language, including voice, sign language, and handwritten text. However, we restrict the scope of this paper to LI of written material that has been digitally encoded.

The ability to identify the language for a written document has a wide range of uses such NLP tasks. It plays an important role in attracting users to a specific website that can provide relevant information for the user's native language (Kralisch and Mandl, 2006). In information retrieval and storage, the procedure of indexing documents in a multilingual collection according to the language they were written in is common. LI is required for document collections where the languages of the documents are unknown at the outset, such as for data retrieved from the World Wide Web (Jauhiainen et al., 2019). It is also suitable for machine translator applications by detecting the user's language without selecting it.

Language Identification is useful also in security. Forensic linguistics is one of the potential applications that use the LI which is considered the link between the legal system and linguistic stylistics (McMenamin, 2002). LI can be used as a methodology for authorship profiling to give proof of an author's linguistic background (Grant, 2007). Authorities and intelligence agencies may be able to learn more about threats and their perpetrators if they can extract more information from an anonymous SMS. Investigators can identify the author of anonymous literature with the use of hints about their languages. In these situations, LI can be used to discover the discriminant language cues in anonymous communication (Abbasi and Chen, 2005).

The study of language acquisition and teaching has received a lot of linguistic attention. The need for resources for language learning has increased because of the growing number of language learners, which has in turn fueled most of the language acquisition research over the past ten years.

The development of the Second Language Acquisition (SLA) theory may potentially benefit from the outcomes of an LI task. The new detection-based approach to transfer articulates the convergence of LI approaches and transfer research (Jarvis, 2010), which was first proposed by Tsur and Rappoport (Tsur and Rappoport, 2007). LI can be used to create teaching strategies, guidelines, and learner feedback that are tailored to each

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 34 - 37

student's mother language. The models specific to each language can be used to create this customized evaluation. For instance, algorithms based on these models could give students feedback in automated writing evaluation systems that is considerably more targeted and concentrated (Rozovskaya and Roth, 2011).

A new word-representation model, Bag of N-Characters (BoNC), has been presented in this work. The proposed model is a derivation of the Bag of Words model (BoW) for characters. Different machine learning algorithms namely; Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Random Forest (RF) and Multi-Layer Perceptron (MLP) have been implemented using BoNC model as a vectorization technique.

The rest of the paper is organized as follows: Section 2 describes the dataset; section 3 describes the system architecture. Experimental settings and results are given in section 4. Finally, the conclusion and future work are presented in section 5.

2 Dataset

The dataset, CoLI-Kenglish, has been distributed by task organizers to the participants (Hosahalli Lakshmaiah et al., 2022). It comprises a set of English and Kannada words written in Roman script. The words are grouped into the following set of classes $\{Kannada, English, Mixed-language, Name, Location, Other\}$

3 System Architecture

The general framework of proposed model consists of three main phases. The first phase is preprocessing where the raw words were prepared to further steps. The second phase is word representation and the third phase is model training. After model construction, the test set was fed to the model for model evaluation. The following are details of each phase.

3.1 Preprocessing

The preprocessing step consists of creating a vocabulary of characters $\mathcal{V}$. In this work, we set the threshold for the occurrence of the character to be considered as $k = 4$.

$$\mathcal{V} = \{ch \mid \textbf{number of occurrence of } ch >= 4\}$$

Figure 1: Word representation vector.

3.2 Word Representation

To represent the training samples, *words*, we used vector of length exceeds the number of characters in the set $\mathcal{V}$ by 2. The components of this vector represents the number of occurrence of the corresponding character as shown in figure 1. The last two components of the vector is reserved for the unknown characters, Kannada characters **<KAN>** and unknown characters **<UNK>**.

3.3 Model Construction

The training samples or words are now represented as vectors. Now, the current phase is model creation. Various machine learning classifiers, namely, support vector machines, random forests and multilayer perceptron have been implemented.

3.3.1 Support Vector Machines

For text classification problems including a significant number of features and documents, as those in the current study. SVM is effective and demonstrated great promise in NLP applications such as dialect identification (Nayel et al., 2021b), rumors detection (Ashraf et al., 2022), sentiment detection (Nayel et al., 2021a), sarcasm detection (Nayel et al., 2021a) and gender biased detection (Elkazzaz et al., 2021).

SVM is a classification technique that generates statistical models that can differentiate between similar classes in the training data. By representing each example in the training data as a point in multidimensional space, it achieves this.

3.3.2 Random Forest (RF)

The random forest is a series of decision trees linked together by several bootstrap samples generated from the original data set. Based on the entropy (or Gini index) of a chosen subset of the features, the nodes are divided. The subsets that are generated using bootstrapping from the original data set, have the same size as the original data set size. Random forests can develop into quite sophisticated predictive models.

3.3.3 Multi-layer Perceptron (MLP)

MLP are adjuncts to feedforward neural networks. It is often used in supervised learning. MLP consists of three types of layers: input layer, output layer and hidden layer. Each layer consists of nodes. The output layer node represents the set of class labels present in the training data set. Learning process in MLP consists of adjusting perceptron weights to make the training data less in errors. This is traditionally done using a back-propagation algorithm that attempts to minimize the MSE.

3.4 Performance Evaluation

We calculated four evaluation metrics, Precision (P), Recall (R), and F1-score to measure the performance of our models. The macro f1-score is the official metric for the shared task (Balouchzahi et al., 2022).

4 Experiments and Results

For preprocessing phase, the threshold is set to be 4, $k = 4$. The vector length was 64, i.e the character vocabulary contains 64 characters. K-folds cross validation technique has been used for the development phase. The training set is divided into three folds, at the first run the first fold has been used as test set and other folds as the training set and so on. Table 1 shows the results of all classifiers for the 3-fold cross validation technique.

Algorithm	Accuracy
RF	64.89%
SGD	65.19%
SVM (Linear)	62.94%
MLP (h=10)	64.49%
MLP (h=20)	64.97%
MLP (h=40)	65.38%

Table 1: 3-fold cross validation accuracy for all classifiers.

Table 2 shows the result of our submission for the shared task for all classifiers. RF proved its superiority and achieved the best performance.

5 Conclusion and Future Work

In this paper, a simple framework for language identification has been introduced. A vectorization approach (BoNC) has been compared. It is clear from the results that RF outperforms all other classifiers. From this study, we can conclude that language identification of text is one of the challenging tasks.

In future work, pre-trained models could be used to improve the performance of classification. Transfer learning can be applied so that knowledge from one domain can be transferred to another domain.

References

A. Abbasi and H. Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

Nsrin Ashraf, Hamada Nayel, and Mohamed Taha. 2022. A comparative study of machine learning approaches for rumors detection in covid-19 tweets. In *2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*, pages 384–387.

Fazlourrahman Balouchzahi, Sabur Butt, Asha Hagde, Noman Ashraf, Shashirekha Hosahalli Lakshmaiah, Grigori Sidorov, and Alexander Gelbukh. 2022. Overview of CoLI-Kanglish: Word Level Language Identification in Code-mixed Kannada-English Texts at ICON 2022. In *19th International Conference on Natural Language Processing Proceedings*.

Fathy Elkazzaz, Fatma Sakr, Rasha Orban, and Hamada Nayel. 2021. BFCAI at ComMA@ICON 2021: Support vector machines for multilingual gender biased and communal language identification. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 70–74, NIT Silchar. NLP Association of India (NLPAI).

Tim Grant. 2007. Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language & the Law*, 14(1).

Shashirekha Hosahalli Lakshmaiah, Fazlourrahman Balouchzahi, Anusha Mudoor Devadas, and Grigori Sidorov. 2022. CoLI-Machine Learning Approaches for Code-mixed Language Identification at the Word Level in Kannada-English Texts. *acta polytechnica hungarica*.

Scott Jarvis. 2010. Comparison-based and detection-based approaches to transfer research. *EUROSLA Yearbook*, 10(1):169–192.

Scott Jarvis and Scott A. Crossley, editors. 2012. *Approaching Language Transfer through Text Classification*. Multilingual Matters, Bristol, Blue Ridge Summit.

	Weighted			Macro F1		
Algorithm	**P**	**R**	**F1-score**	**P**	**R**	**F1-score**
RF	0.73	0.73	0.72	0.52	0.41	0.43
SGD	0.74	0.73	0.72	0.51	0.43	0.41
SVM	0.73	0.73	0.72	0.42	0.36	0.36
MLP($h = 20$)	0.73	0.72	0.70	0.43	0.34	0.34
MLP($h = 10$)	0.72	0.72	0.70	0.42	0.34	0.34

Table 2: Results of our submissions on test set

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *J. Artif. Int. Res.*, 65(1):675–682.

A. Kralisch and T. Mandl. 2006. Barriers to information access across languages on the internet: Network and language effects. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, volume 3, pages 54b–54b.

Haizhou Li, Bin Ma, and Kong Aik Lee. 2013. Spoken language recognition: From fundamentals to practice. *Proceedings of the IEEE*, 101(5):1136–1159.

Gerald R McMenamin. 2002. *Forensic linguistics: Advances in forensic stylistics*. CRC press.

Hamada Nayel, Eslam Amer, Aya Allam, and Hanya Abdallah. 2021a. Machine learning-based model for sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 386–389, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Hamada Nayel, Ahmed Hassan, Mahmoud Sobhi, and Ahmed El-Sawy. 2021b. Machine learning-based approach for Arabic dialect identification. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 287–290, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Alla Rozovskaya and Dan Roth. 2011. Algorithm selection and model adaptation for ESL correction tasks. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 924–933, Portland, Oregon, USA. Association for Computational Linguistics.

Laura Mayfield Tomokiyo and Rosie Jones. 2001. You're not from 'round here, are you? naive Bayes detection of non-native utterances. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16, Prague, Czech Republic. Association for Computational Linguistics.